Bernoulli Distribution
$X \sim Ber(p)$ if it takes on only two values: 0 or 1, with probabilities $1-p$ and p, respectively. Its PMF is

$$P(X=x) = \begin{cases} p^x(1-p)^x & \text{if } x = 0 \text{ or } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

Its MGF, $M(t) = 1 - p + pe^t$. The mean and variance are p and $p(1-p)$ respectively.

Binomial Distribution
Let $X_1, \ldots, X_n$ be sequence of i.i.d Bernoulli trials. Then, the no. of successes amongst the first n trials is given by $Y = X_1 + \cdots + X_n$ i.e. $Y \sim Bin(n,p)$

$$P(Y=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The MGF is $(1-p+pe^t)^n$. The mean and variance are np and $np(1-p)$ respectively.

Poisson Distribution
The poisson distribution is defined over the parameter $\lambda > 0$. Its PMF is

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$E[X] = Var(X) = \lambda$, and MGF $M(t) = e^{\lambda(e^t-1)}$.

Geometric Distribution
$X \sim Geom(p)$ interpreted as the number of trials until the first success.

$$P(X=k) = (1-p)^{k-1}p$$

$E[X] = 1/p$ and $Var(X) = (1-p)/p^2$, and MGF, $M(t) = pe^t/(1-(1-p)e^t)$.

Note: (Tail Probability Formula)

$$P(X \geq K) = \sum_{i=k}^{\infty} p(1-p)^{i-1} = p\frac{(1-p)^{k-1}}{1-(1-p)} = (1-p)^{k-1}$$

Negative Binomial Distribution
Suppose that a sequence of independent trials performed such that there are r successes in total. i.e. $X \sim NegBin(r,p)$

$$P(X=k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r$$

Equivalently, $X = Y_1 + \cdots + Y_r$ where $Y_i's$ are i.i.d. $Geom(p)$. $E[X] = r/p$ and $Var(x) = r(1-p)/p^2$

Hypergeometric Distribution
Given population of size N, of which M are of type I, $N-M$ are of type II, want to draw n samples without replacement, then

$$P(X=k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \qquad 0 \leq k \leq n$$

where X is the number of Type I selected.

Uniform Random Variable
$X \sim Unif(a,b)$ where $a < b$ if

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Note that for any $(x,y) \in [a,b]$:

$$P(x < X < y) = \frac{y-x}{b-a}$$

Mean, Variance and MGF:

$$\mathrm{E}[X] = \frac{a+b}{2}, \mathrm{Var}(X) = \frac{(b-a)^2}{12}, M(t) = \frac{e^t-1}{t}$$

CDF:

$$F_X(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b] \\ 1 & \text{if } x > b \end{cases}$$

Exponential Random Variable
$X \sim Exp(\lambda)$ where $\lambda > 0$ if it has the pdf

$$f_X(x) = \lambda e^{-\lambda x} \quad x \geq 0 \quad (0 \text{ otherwise})$$

whereby its mean and variance are given by $1/\lambda$ and $1/\lambda^2$ respectively. MGF is $\lambda/(\lambda - t)$.

Note: (Tail Probability Formula)

$$P(x > t) = \int_t^\infty \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_t^\infty = e^{-\lambda t}$$

Gamma Distribution
The gamma distribution is defined over two parameters, $\alpha > 0$, $\lambda > 0$. Its PDF is

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \quad x \geq 0$$

The MGF is $\left(1 - \frac{t}{\lambda}\right)^{-\alpha}$ for $t < \lambda$. The mean and variance are α/λ and α/λ^2 respectively.

Note: Gamma is the generalization of Exponential
Let $X = Y_1 + \cdots + Y_\alpha$ where Y_i's are i.i.d $Exp(\lambda)$. Then, $X \sim \Gamma(\alpha, \lambda)$.

Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

The MGF is $e^{\mu t + \sigma^2 t^2/2}$. Its mean and variance are μ and σ^2 respectively. For the standard normal, its CDF is given by

$$F_Z = \int_\infty \frac{1}{\sqrt{2\pi}} \exp -x^2/2$$

Functions of a Random Variable
Suppose X has density function $f(x)$. We want to find the density of $Y = g(X)$ for some given function $g(x)$.

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P\left(X \leq g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right)$$

assuming $g(x)$ is a differentiable, strictly increasing.

Proposition: Let $U \sim Unif[0,1]$, and let $X = F^{-1}(U)$. Then the CDF of X is F.

Sums of Independent Random Variables
Suppose we want to find the distribution of $\phi(X,Y) = X + Y$ for independent X and Y.

$$f_Z(z) = \sum_y f_X(z-y)f_Y(y) = \sum_x f_Y(z-x)f_X(x)$$

$$f_Z(s) = \int f_X(z-y)f_Y(y)dy = \int f_Y(z-x)f_X(x)dx$$

Multidimensional Change of Variables
Suppose that X and Y are jointly distributed continuous random variables, that X and Y are mapped onto U and V by the transformation

$$u = g_1(x,y) \quad v = g_2(x,y)$$

and that the transformation can be inverted to obtain

$$x = h_1(u,v) \quad y = h_2(u,v)$$

Assume that g_1 and g_2 have continuous partial derivatives and that the Jacobian

$$J(x,y) = \det\begin{pmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{pmatrix} \neq 0$$

for all x and y. The joint density of U and V is

$$f_{UV}(u,v) = f_{XY}(h_1(u,v), h_2(u,v))\left|J^{-1}(h_1(u,v), h_2(u,v))\right|$$

Conditional Distributions
Conditional distribution of X given Y is defined by

$$f_{X|Y}(x|y) = \frac{f_{X,Y(x,y)}}{f_Y(y)}$$

Independent Random Variables
X and Y are independent if the conditional distribution of X given $Y = y$ does not depend on y, i.e. $f_{X|Y}(\cdot|y) = f_X(\cdot)$. Equivalently,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

Extrema and Order Statistics
Let U denote the maximum of the X_i's and V the minimum, where X_i's are independent. Then,

$$\begin{aligned} F_U(u) &= P(U \leq u) = P(X_1 \leq u) \ldots P(X_n \leq u) \\ &= [F(u)]^n \end{aligned}$$

$$f_U(u) = \frac{d}{du}F_U(u) = nf(u)[F(u)]^{n-1}$$

$$\begin{aligned} F_V(v) &= 1 - P(V > v) \\ &= 1 - P(X_1 > v) \ldots P(X_n > v) \\ &= 1 - [1 - F(v)]^n \end{aligned}$$

$$f_V(v) = \frac{d}{dv}F_V(v) = nf(v)[1-F(v)]^{n-1}$$

Let $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ denote the order statistics. Thus, $X_{(n)}$ is the maximum, and $X_{(1)}$ is the minimum.

The density of the k th-order statistic, $X_{(k)}$ is:

$$f_k(x) = \frac{n!}{(k-1)!(n-k)!}f(x)[F(x)]^{k-1}[1-F(x)]^{n-k}$$

Joint Density of $V = X_{(1)}$ and $U = X_{(n)}$:

$$f(u,v) = n(n-1)f(v)f(u)[F(u) - F(v)]^{n-2}, \quad u \geq v$$

Expectation

$$\mathrm{E}[X] = \sum_i x_i p(x_i) \quad \text{(discrete)}$$

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} xf(x)\ dx \quad \text{(continuous)}$$

Markov's Inequality
Suppose X non-negative such that $\mathrm{E}[X] < \infty$, then for any $t > 0$,

$$P(X \geq t) \leq \frac{\mathrm{E}[X]}{t}$$

Expectation of Functions of Multiple RVs
Expectations of functions of (X,Y) are defined by

$$\mathrm{E}[\phi(X,Y)] = \sum_{x,y}\phi(x,y)f_{X,Y}(x,y)$$

$$\mathrm{E}[\phi(X,Y)] = \int\int\phi(x,y)f_{X,Y}(x,y)dxdy$$

Variance

$$\begin{aligned} Var(X) &= \mathrm{E}[(X - \mathrm{E}[X])^2] \\ &= \mathrm{E}[X^2] - \mathrm{E}[X]^2 \geq 0 \end{aligned}$$

And

$$Var(a + bX) = b^2 Var(X)$$

Chebyshev's Inequality
Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$,

$$P(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}$$

i.e. using the moment of X to bound the distribution of X.

Covariance

$$\begin{aligned} Cov(X,Y) &= \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] \end{aligned}$$

Corollary:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2Cov(X,Y)$$

$$Var(\sum_{i=1}^n Xi) = \sum_{i=1}^n Var(X_i) \quad \text{if } X_i\text{'s independent}$$

Correlation

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Conditional Expectation

$$E(Y|X=x) = \sum_y yp_{Y|X}(y|x) \quad \text{discrete}$$

$$E(Y|X=x) = \int_y yf_{Y|X}(y|x)dy \quad \text{continuous}$$

Conditional Variance

$$Var(Y|X) = \mathrm{E}[(Y - E(Y))^2|X] = E(Y^2|X) - [E(Y|X)]^2$$

Law of Total Expectation

$$E(Y) = \mathrm{E}[E(Y|X)]$$

Equivalently,

$$E(Y) = \sum_x E(Y|X=x) p_X(x) \quad \text{discrete}$$

$$E(Y) = \int_x E(Y|X=x) f_X(x) dx \quad \text{continuous}$$

Law of Total Variance

$$Var(Y) = Var[E(Y|X)] + \mathrm{E}[Var(Y|X)]$$

Moment Generating Functions

The moment generating function (MGF) of a random variable X is,

$$M(t) = \mathrm{E}[e^{tX}]$$

and the r^{th} moment of a random variable is $\mathrm{E}[X^r]$ if it exists. MGF uniquely determines the distribution.

Property: $M^{(r)}(0) = \mathrm{E}[X^r] = \frac{d^r M(t)}{d^r t}|_t = 0$
Property: If X has the mgf $M_X(t)$ and $Y = a + bX$, then Y has the mgf $M_Y(t) = e^{at} M_X(bt)$
Multiplicative Property: If X and Y independent, $Z = X + Y$, then $M_Z(t) = M_X(t) M_Y(t)$

Chernoff Bounds

$$P(X \geq a) \leq e^{-ta} M_X(t) \quad \text{for all } t > 0$$

$$P(X \leq a) \leq e^{-ta} M_X(t) \quad \text{for all } t < 0$$

Jensen's Inequality

If $f(x)$ is a convex function, then

$$\mathrm{E}[f(X)] \geq f(E(X))$$

provided the expectation exists and is finite.
Note: $f(x)$ said to be convex if $f''(x) \geq 0$ for all x.

Weak Law of Large Numbers

Let $X_1, X_2, \ldots, X_i \ldots$ be i.i.d sequence with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then, for any $\epsilon > 0$,

$$P\left(|\bar{X}_n - \mu| > \epsilon\right) \to 0 \text{ as } n \to \infty$$

i.e. convergence in probability

Continuity Theorem

If $M_n(t) \to M(t)$ for all t in an open interval containing zero, then $F_n(x) \to F(x)$ at all continuity points of F. i.e. MGFs are useful to prove convergence in distribution.

Central Limit Theorem

Let $X_1, X_2, \ldots$ be i.i.d sequence of RVs with mean μ and variance σ^2 and the common distribution function F. Let $S_n = \sum_{i=1}^n X_i$
Then

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x), \quad -\infty < x < \infty$$

Method of Moments Estimators

The k-th moment of a probability law is defined as

$$\mu_k = \mathrm{E}[X^k]$$

If $X_1, X_2, \ldots, X_n$ are iid random variables from that distribution, the k-th sample moment is defined as

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

where $\hat{\mu}_k$ can be taken as an estimate of μ_k.

Maximum Likelihood Estimators

Objective is to maximize the likelihood function

$$L(\theta) = \prod_{i=1}^n f(x_i|\theta)$$

Alternatively, maximize log-likelihood function:

$$l(\theta) = \sum_{i=1}^n \log f(x_i|\theta)$$

MLE for Multinomial Cell Probabilities

The likelihood is the joint frequency function

$$\mathrm{lik}(p_1, \ldots, p_m) = \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i}$$

Note: Marginal distribution of each X_i is $Bin(n, p_i)$. The log likelihood is

$$l(p_1, \ldots, p_m) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i$$

Bootstrap Procedure - Multinomial Example

Boostrap procedure for approximating the sampling distributions of $\hat{\theta}$:

1. Assume multinomial distribution with $\hat{\theta}$ provides a good fit to the data
2. Simulate N random samples from multinomial distribution with corresponding probabilities p_1, p_2, p_3 and $n = 1029$.
3. For ea. random sample, calculate MLE, θ^* of θ
4. Use the N values of θ^* to approximate the sampling distributions of $\hat{\theta}$

The standard error of $\hat{\theta}$ can be estimated using

$$s_{\hat{\theta}} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\theta_i^* - \bar{\theta})^2} \quad \text{where} \quad \bar{\theta} = \frac{1}{N} \sum_{i=1}^N \theta_i^*$$

Consistency

Let $\hat{\theta}_n$ be an estimate of a parameter θ_0 based on a sample of size n. $\hat{\theta}_n$ is said to be consistent in probability if $\hat{\theta}_n$ converges in probability to θ_0 as n approaches infinity. That is, for $\epsilon > 0$,

$$P(|\hat{\theta}_n - \theta_0| > \epsilon) \to 0 \text{ as } n \to \infty$$

Fisher Information (Lemma A)

$$I(\theta) = \mathrm{E}\left[\frac{\partial}{\partial\theta} \log f(X|\theta)\right]^2$$

$$= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2} \log f(X|\theta)\right]$$

Large Sample Theory for MLE

Let $\hat{\theta}$ denote the MLE of θ_0. The probability distribution of

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$$

tends to a standard normal distribution. Therefore, the asymptotic variance of the MLE is

$$\frac{1}{nI(\theta)} = -\frac{1}{\mathrm{E}[l''(\theta_0)]}$$

Approximate Confidence Intervals

Confidence intervals can be approximated through the large sample theory for MLE by taking $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \to N(0,1)$, as $n \to \infty$.

$$P\left(-z(\alpha/2) \leq \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) \leq z(\alpha/2)\right) \approx 1 - \alpha$$

An approximate large sample $(1-\alpha)100\%$ confidence interval for θ_0 is

$$\hat{\theta} \pm z(\alpha/2) \frac{1}{\sqrt{nI(\hat{\theta})}}$$

Bootstrap Confidence Interval

Suppose that $\underline{\theta}$ and $\bar{\theta}$ are lower and upper quantiles of the distribution of θ^*. Let $\underline{\delta} = \underline{\theta} - \hat{\theta}$ and $\bar{\delta} = \bar{\theta} - \hat{\theta}$, then the approximate $(1-\alpha)100\%$ CI is given by

$$(\hat{\theta} - \bar{\delta}, \hat{\theta} - \underline{\delta})$$

Efficiency

The efficiency of two estimators, $\hat{\theta}_0, \hat{\theta}_1$ is given as

$$\mathrm{eff}(\hat{\theta}_0, \hat{\theta}_1) := \mathrm{Var}(\hat{\theta}_1) / \mathrm{Var}(\hat{\theta}_0)$$

If the effficiency is smaller than 1, then $\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_0)$

Cramer-Rao Lower Bound

Under smoothness assumptions of a $f(x|\theta)$ for a statistic $T := t(X_1, \cdots, X_n)$

$$\mathrm{Var}(T) \geq \frac{1}{nI(\theta)}$$

This gives the lower bound for the variance of any estimator of θ.

Sufficiency

A statistic $T(X_1, \cdots, X_n)$ is said to be sufficient for θ if the conditional distribution of $X_1, \cdots, X_n$ given $T = t$ does not depend on θ for any value of t. If T is sufficient for θ, the MLE for θ is a function only of T.

Factorization Theorem

The statistic $T(X_1, \cdots, X_n)$ is sufficient for a parameter θ iff the joint pdf factorises in the form

$$f(\vec{x}|\theta) = g(T(\vec{x}), \theta) h(\vec{X})$$

Exponential Family of Probability Distributions

1-parameter members of the exponential family have pdfs or pmfs in the form

$$f(x|\theta) = \begin{cases} \exp\{c(\theta)T(x) + d(\theta) + S(x)\}, & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

where the set A does not depend on θ.
Suppose that $X_1, X_2, \ldots, X_n$ i.i.d. with joint pdf

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^n \exp\left[c(\theta) T(x_i) + d(\theta) + S(x_i)\right]$$

$$= \exp\left[c(\theta) \sum_{i=1}^n T(x_i) + nd(\theta)\right]$$

$$\times \exp\left[\sum_{i=1}^n S(x_i)\right]$$

$$= g\left(\sum_{i=1}^n T(x_i), \theta\right) \times h(x_1, x_2, \ldots, x_n)$$

Then, $\sum_{i=1}^n T(x_i)$ is a sufficient statistic for θ.

Rao-Blackwell Theorem

Let $\hat{\theta}$ be an estimator of θ with $E(\hat{\theta})^2 < \infty$. Suppose that T is sufficient for θ and let $\tilde{\theta} = E(\hat{\theta}|T)$. Then, for all θ,

$$\mathrm{E}[(\tilde{\theta} - \theta)^2] \leq \mathrm{E}[(\hat{\theta} - \theta)^2]$$

Delta Method

Given knowledge of μ_X and σ_X^2, but not the underlying distribution, want to find the mean and variance of $Y = g(X)$. First-Order Approximation:

$$Y = g(X) \approx g(\mu_X) + (X - \mu_X) g'(\mu_X)$$

$$\mu_Y = E(g(X)) \approx g(\mu_X) \quad \text{and} \quad \sigma_Y^2 \approx \sigma_X^2 \left[g'(\mu_X)\right]^2$$

Second-Order Approximation

$$g(X) \approx g(\mu_X) + (X - \mu_X) g'(\mu_X) + \frac{1}{2}(X - \mu_X)^2 g''(\mu_X)$$

$$\mu_Y = E(g(X)) \approx g(\mu_X) + \frac{1}{2}\sigma_X^2 g''(\mu_X)$$

Bayesian Inference

Let unknown parameter Θ be a random variable with prior distribution. The posterior distribution is given by:

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{X,\Theta}(x, \theta)}{f_X(x)}$$

$$= \frac{f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)}{\int f_{X|\Theta}(x \mid \theta) f_\Theta(\theta) d\theta}$$

i.e. Posterior density $\propto$ Likelihood $\times$ Prior density

Useful Results: Beta Distribution

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 \leq x \leq 1$$

$$E(X) = \frac{a}{a+b} \quad , \quad Var(X) = \frac{ab}{(a+b)^2(a+b+)}$$

Beta Integral

$$\int_0^1 x^{a-1}(1-x)^{b-1} dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

where

$$\Gamma(n) = (n-1))! \quad , \quad \Gamma(n+1) = n\Gamma(n)$$

Other Stuff

$$\int_0^\infty ye^{-y} dy = 1$$

$$\int_0^\infty y^2 e^{-y} dy = \left(y^2(-e^{-y})|_0^\infty + \int_0^\infty 2ye^{-y} dy\right)$$

$$= 0 + 2\int_0^\infty ye^{-y} dy$$

$$\sum_{k=m}^\infty ar^k = \frac{ar^m}{1-r} \text{ where } |r| < 1 \text{ , } e^x = \sum_{n=0}^\infty \frac{x^n}{n!}$$

$$\ln(1-x) = -\sum_{n=1}^\infty \frac{x^n}{n} \text{ , } \ln(1+x) = \sum_{n=1}^\infty (-1)^{n+1} \frac{x^n}{n}$$